Could There be a Turing Test for Qualia?

[Published in Müller, Vincent C. and Ayesh, Aladdin (eds.) (2012), *Revisiting Turing and his Test: Comprehensiveness, Qualia, and the Real World (AISB/IACAP Symposium)* (Hove: AISB), pp. 41-48, ISBN 978-1-908187-15-4]

Could There be a Turing Test for Qualia?

Paul Schweizer[1]

Abstract. The paper examines the possibility of a Turing test designed to answer the question of whether a computational artefact is a genuine subject of conscious experience. Even given the severe epistemological difficulties surrounding the 'other minds problem' in philosophy, we nonetheless generally believe that other human beings are conscious. Hence Turing attempts to defend his original test (2T) in terms of operational parity with the evidence at our disposal in the case of attributing understanding and consciousness to other humans. Following this same line of reasoning, I argue that the conversation-based 2T is far too weak, and we must scale up to the full linguistic and robotic standards of the Total Turing Test (3T).

Within this framework, I deploy Block's distinction between Phenomenal-consciousness and Access-consciousness to argue that passing the 3T could at most provide a sufficient condition for concluding that the robot enjoys the latter but not the former. However, I then propose a variation on the 3T, adopting Dennett's method of 'heterophenomenology', to rigorously probe the robot's purported 'inner' qualitative experiences. If the robot could pass such a prolonged and intensive Qualia 3T (Q3T), then the purely behavioural evidence *would* seem to attain genuine parity with the human case. Although success at the Q3T would not supply definitive proof that the robot was genuinely a subject of Phenomenal-consciousness, given that the external evidence is now equivalent with the human case, apparently the only grounds for denying qualia would be appeal to difference of *internal* structure, either physical-physiological or functional-computational. In turn, both of these avenues are briefly examined.

1 INTRODUCTION

According to the widely embraced 'computational paradigm', which underpins cognitive science, Strong AI and various allied positions in the philosophy of mind, computation (of one sort or another) is held to provide the scientific key to explaining mentality in general and, ultimately, to reproducing it artificially. The paradigm maintains that cognitive processes are essentially computational processes, and hence that intelligence in the natural world arises when a material system implements the appropriate kind of computational formalism. So this broadly Computational Theory of Mind (CTM) holds that the mental states, properties and contents sustained by human beings are fundamentally computational in nature, and that computation, at least in principle, opens the possibility of creating artificial minds with comparable states, properties and contents.

Traditionally there are two basic features that are held to be essential to minds and which decisively distinguish mental from non-mental systems. One is representational content: mental states can be *about* external objects and states of affairs. The other is conscious experience: roughly and as a first approximation, there is *something it is like* to be a mind, to be a particular mental subject. As a case in point, there is something it is like for me to be consciously aware of typing this text into my desk top computer. Additionally, various states of my mind are concurrently directed towards a number of different external objects and states of affairs, such as the letters that appear on my monitor. In stark contrast, the table supporting my desk top computer is not a mental system: there are no states of the table that are properly about anything, and there is nothing it is like to be the table.

Just as it seems doubtful that the term 'mind' should be applied to a system with no representational states, so too, many would claim that a system entirely devoid of conscious experience cannot be a mind. Hence if the project of Strong AI is to be successful at its ultimate goal of producing a system that truly counts as an artificially engendered locus of mentality, then it would seem necessary that this computational artefact be fully conscious in a manner comparable to human beings.

2 CONSCIOUSNESS AND THE ORIGINAL TURING TEST

In 1950 Turing [1] famously proposed an answer to the question 'Can (or could) a machine think?' by replacing it with the more precise and empirically tractable question 'Can (or could) a machine pass a certain type of test?', which mode of assessment has since become universally referred to as the 'Turing test' (2T). In brief, (the standardized version of) Turing's test is an 'imitation game' involving three players: a computational artifact and two humans. One of the humans is the 'judge' and can pose questions to the remaining two players, where the goal of the game is for the questioner to determine which of the two respondents is the computer. If, after a set amount of time, the questioner guesses correctly, then the machine loses the game, and if the questioner is wrong then the machine wins. Turing claimed, as a basic theoretical point, that any machine that could win the game a suitable number of times has passed the test and should be judged to be intelligent, in the sense that its behavioral performance has been demonstrated to be indistinguishable from that of a human being.

[1] Institute for Language, Cognition and Computation, School of Informatics, Univ. of Edinburgh, EH8 9AD, UK. Email: paul@inf.ed.ac.uk.

In his prescient and ground breaking article, Turing explicitly considers the application of his test to the question of *machine consciousness*. This is in section (4) of the paper, where he considers the anticipated 'Argument from Consciousness' objection to the validity of his proposed standard for answering the question 'Can a machine think?'. The objection is that, as per the above, consciousness is a necessary precondition for genuine thinking and mentality, and that a machine might fool its interlocutor and pass the purely behavioural 2T, and yet remain completely devoid of internal conscious experience. Hence merely passing the 2T does not provide a sufficient condition for concluding that the system in question possesses the characteristics required for intelligence and *bona fide* thinking. Hence the 2T is inherently defective.

Turing's defensive strategy is to invoke the well known and severe epistemological difficulties surrounding the very same question regarding our fellow human beings. This is the notorious 'other minds problem' in philosophy – how do you know that other people actually have a conscious inner life like your own? Perhaps everyone else is a zombie and you're the only conscious being in the universe. As Turing humorously notes, this type of 'solipsistic' view (although more accurately characterized as a form of other minds skepticism, rather than full blown solipsism), while logically impeccable, tends to make communication difficult, and rather than continually arguing over the point, it is usual to simply adopt the polite convention that everyone is conscious.

Turing notes that on its most extreme construal, the only way that one could be sure that a machine or another human being is conscious and hence genuinely thinking is to *be* the machine or the human and *feel oneself* thinking. In other words, one would have to gain first person access to *what it's like* to be the agent in question. And since this is not an empirical option, we can't know with certainty whether any other system is conscious – all we have to go on is behaviour. Hence Turing attempts to justify his behavioural test that a machine can think, and *ipso facto*, has conscious experience, by claiming parity with the evidence at our disposal in the case of other humans. He therefore presents his anticipated objector with the following dichotomy: either be guilty of an inconsistency by accepting the behavioural standard in the case of humans but not computers, or maintain consistency by rejecting it in *both* cases and embracing solipsism. He concludes that most consistent proponents of the argument from consciousness would chose to abandon their objection and accept his test rather than be forced into the solipsistic position.

However, it is worth applying some critical scrutiny to Turing's reasoning at this early juncture. Basically, he seems to be running *epistemological* issues together with *semantical* and/or *factive* questions which should properly be kept separate. It's one thing to ask what we *mean* by saying that a system has a mind – i.e. what essential traits and properties are we ascribing to it with the use of the term; while it's quite another thing to ask how we can *know* that a given system actually satisfies this meaning and hence really does have a mind. Turing's behaviouristic methodology has a strong tendency to collapse these two themes, but it is important to note that they are conceptually distinct. In the argument from consciousness, the point is that we *mean* something substantive, something more than just verbal stimulus-response patterns, when we attribute mentality to a system. In this case the claim is that we mean that the system in question has conscious experience, and this property is required for any agent to be accurately described with the term 'mind'.

So one could potentially hold that consciousness is essential to mentality (because that's part of the core meaning of the term) and that:

(1) other human beings are in fact conscious
(2) the computer is in fact unconscious
(3) therefore, the computer doesn't have a mind, even though it passes the 2T.

This could be the objective state of affairs that genuinely obtains in the world, and this is completely independent of whether we can *know*, with certainty, that premises (1) and (2) are actually true. Although epistemological and factive issues are intimately related and together inform our general practices and goals of inquiry, nonetheless we could still be correct in our assertion, without being able to *prove* it's correctness. So if one thought that consciousness was essential to genuine mentality, then one could seemingly deny that any purely behaviouristic standard was sufficient to test for whether a system had or was a mind.

In the case of other human beings, we certainly take behaviour as *evidence* that they are conscious, but the evidence could in principle overwhelmingly support a *false* conclusion, in both directions. For example, someone could be in a comatose state where they could show no evidence of being conscious because they could make no bodily responses. But in itself this wouldn't make them unconscious. They could still be cognizant of what was going on and perhaps be able to report, retrospectively, on past events once out of their coma. And again, maybe *some* people really are zombies, or sleepwalkers, and exhibit all the appropriate external signs of consciousness even though they're really asleep or under some voodoo spell - it's certainly a conceivable state of affairs which cannot simply be ruled out *a priori*.

Historically, there has been disagreement regarding the proper interpretation of Turing's position regarding the intended import of his test. Some have claimed that the 2T is proposed as an operational *definition* of intelligence, thinking, etc., (e.g. Block [2], French [3]), and as such it has immediate and fundamental faults. However, in the current discussion I will adopt a weaker reading and interpret the test as purporting to furnish an empirically specifiable criterion for when intelligence can be legitimately *ascribed* to an artefact. On this reading, the main role of behavior is inductive or evidential rather than constitutive, and so behavioral tests for mentality do not provide a necessary condition nor a reductive definition. At most, all that is warranted is a *positive* ascription of intelligence or mentality, *if* the test is adequate *and* the system passes. In the case of Turing's 1950 proposal, the adequacy of the test is defended almost entirely in terms of parity of input/output performance with human beings, and hence alleges to employ the same operational standards that we tacitly adopt when ascribing conscious thought processes to our fellow creatures.

Thus the issue would appear to hinge upon the degree of evidence a successful 2T performance provides for a positive conclusion in the case of a computational artefact, (i.e. for the negation of (2) above), and how this compares to the total body of evidence that we have in support of our belief in the truth of (1). We will only be guilty of an inconsistency or employing a double standard if the two are on a par and we nonetheless dogmatically still insist on the truth of both (1) and (2). But if it

turns out to be the case that our evidence for (1) is significantly better than for the negation of (2), then we are not forced into Turing's dichotomy. And in terms of the original 2T, I think there is clearly very little parity with the human case. We rely on far more than simply *verbal* behaviour in arriving at the polite convention that other human beings are conscious. In addition to conversational data, we lean very heavily on their bodily actions involving perception of the spatial environment, navigation, physical interaction, verbal and other modes of response to communally accessible non-verbal stimuli in the shared physical surroundings, etc. So the purely conversational standards of the 2T are not nearly enough to support a claim of operational parity with humans. In light of the foregoing observations, in order to move towards evidential equivalence in terms of observable behaviour, it is necessary to break out of the closed syntactic bubble of the 2T and scale up to a full linguistic *and robotic* version of the test. But before exploring this vastly strengthened variation as a potential test for the presence of conscious experience in computational artefacts, in the next section I will briefly examine the notion of consciousness itself, since we first need to attain some clarification regarding the phenomenon in question, before we go looking for it in robots.

3 TWO TYPES OF CONSCIOUSNESS

Even in the familiar human case, consciousness is a notoriously elusive phenomenon, and is quite difficult to characterize rigorously. In addition, the word 'consciousness' is not used in a uniform and univocal manner, but rather appears to have different meanings in different contexts of use and across diverse academic communities. Block [4] provides a potentially illuminating philosophical analysis of the distinction and possible relationship between two common uses of the word. Block contends that consciousness is a 'mongrel' term connoting a number of different concepts and denoting a number of different phenomena. He attempts to clarify the issue by distinguishing two basic and distinct forms of consciousness that are often conflated: *Phenomenal* or P-consciousness and *Access* or A-consciousness. Very roughly, "Phenomenal consciousness is experience: what makes a state phenomenally conscious is that there is 'something it's like' to be in that state". Somewhat more controversially, Block holds that P-conscious properties, as such, are "distinct from any cognitive, intentional or functional property." The notoriously difficult explanatory gap problem in philosophical theorizing concerns P-consciousness – e.g. how is it possible that appeal to a physical brain process could explain what it is like to see something as red?

So we must take care to distinguish this type of purely qualitative, Phenomenal consciousness, from Access consciousness, the latter of which Block sees as an *information processing* correlate of P-consciousness. A-consciousness states and structures are those which are directly available for control of speech, reasoning and action. Hence Block's rendition of A-consciousness is similar to Baars' [5] notion that conscious *representations* are those that are broadcast in a global workspace. The functional/computational approach holds that the level of analysis relevant for understanding the mind is one that allows for multiple realization, so that in principle the same mental states and phenomena can occur in vastly different types of physical systems which implement the same abstract functional or computational structure. As a consequence, a staunch adherent of the functional-computational approach is committed to the view that the same *conscious* states must be preserved across widely diverse type of physical implementation. In contrast, a more 'biological' approach holds that details of the particular physical/physiological realization matter in the case of conscious states. Block says that if P = A, then the information processing side is right, while if the biological nature of experience is crucial then we can expect that P and A will diverge.

A crude difference between the two in terms of overall characterization is that P-consciousness content is qualitative while A-consciousness content is representational. A-conscious states are necessarily transitive or intentionally directed, they are always states of consciousness *of*. However. P-conscious states don't have to be transitive. On Block's account, the paradigm P-conscious states are the qualia associated with sensations, while the paradigm A-conscious states are propositional attitudes. He maintains that the A-type is nonetheless a genuine form of consciousness, and tends to be what people in cognitive neuroscience have in mind, while philosophers are traditionally more concerned with qualia and P-consciousness, as in the hard problem and the explanatory gap. In turn, this difference in meaning can lead to mutual misunderstanding. In the following discussion I will examine the consequences of the distinction between these two types of consciousness on the prospects of a Turing test for consciousness in artefacts.

4 THE TOTAL TURING TEST

In order to attain operational parity with the evidence at our command in the case of human beings, a Turing test for even basic linguistic understanding and intelligence, let alone conscious experience, must go far beyond Turing's original proposal. The conversational 2T relies solely on verbal input/output patterns, and these alone are not sufficient to evince a correct *interpretation* of the manipulated strings. Language is primarily about *extra-linguistic* entities and states of affairs, and there is nothing in a cunningly designed program for pure syntax manipulation which allows it to break free of this closed loop of symbols and demonstrate a proper correlation between word and object. When it comes to judging human language users in normal contexts, we rely on a far richer domain of evidence. Even when the primary focus of investigation is language proficiency and comprehension, sheer *linguistic* input/output data is not enough. Turing's original test is not a sufficient condition for concluding that the computer genuinely understands or refers to anything with the strings of symbols it produces, because the computer doesn't have the right sort of relations and interactions with the objects and states of affairs *in the real world* that its words are supposed to be about. To illustrate the point; if the computer has no eyes, no hands, no mouth, and has never seen or eaten anything, then it is not talking about hamburgers when its program generates the string of English symbols 'h-a-m-b-u-r-g-e-r-s' – it's merely operating inside a closed loop of syntax.

In sharp contrast, *our* talk of hamburgers is intimately connected to *nonverbal* transactions with the objects of reference. There are 'language entry rules' taking us from nonverbal stimuli to appropriate linguistic behaviours. When given the visual stimulus of being presented with a pizza, a taco and a kebab, we can produce the salient utterance "Those

particular foodstuffs are not hamburgers". And there are 'language exit rules' taking us from linguistic expressions to appropriate nonverbal actions. For example, we can follow complex verbal instructions and produce the indicated patterns of behaviour, such as finding the nearest Burger King on the basis of a description of its location in spoken English. Mastery of both of these types of rules is essential for deeming that a human agent understands natural language and is using expressions in a correct and referential manner - and the hapless 2T computer *lacks* both.[2]

And when it comes to testing for conscious experience, we again need these basic additional dimensions of perception and action *in the real world* as an essential precondition. The fundamental limitations of mere conversational performance naturally suggest a strengthening of the 2T, later named the Total Turing Test (3T) by Harnad [7], wherein the repertoire of relevant behaviour is expanded to include the full range of intelligent human activities. This will require that the computational procedures respond to and control not simply a teletype system for written inputs and outputs, but rather a well crafted artificial body. Thus in the 3T the scrutinized artefact is a *robot*, and the data to be tested coincide with the full spectrum of behaviours of which human beings are normally capable. In order to succeed, the 3T candidate must be able to do, in the real world of objects and people, everything that intelligent people can do. Thus Harnad expresses a widely held view when he claims that the 3T is "...no less (nor more) exacting a test of having a mind than the means we already use with one another... [and, echoing Turing] there is no stronger test, short of *being* the candidate". And, as noted above, the latter state of affairs is not an empirical option. examined.[3]

Since the 3T requires the ability to perceive and act in the real world, and since A-consciousness states and structures are those which are directly available for control of speech, reasoning and action, it would seem to follow that the successful 3T robot must be A-conscious. For example, in order to pass the test, the robot would have to behave in an appropriate manner in any number of different scenarios such as the following. The robot is handed a silver platter on which a banana, a boiled egg, a teapot and a hamburger are laid out. The robot is asked to pick up the piece of fruit and throw it out the window. Clearly the robot could not perform the indicated action unless it had direct information processing access to the identity *of* the salient object, its spatial location, the movements *of* its own mechanical arm, the location and geometrical properties *of* the window, etc. Such transitive, intentionally directed A-conscious states are plainly required for the robot to pass the test.

But does it follow that the successful 3T robot is P-conscious? It seems, not, since on the face of it there appears to be no reason why the robot could not pass the test relying on A-consciousness alone. All that is being tested is its executive control of the cognitive processes enabling it to reason correctly and perform appropriate verbal and bodily actions in response to a myriad of linguistic and perceptual inputs. These abilities are demonstrated solely through its external behaviour, and so far, there seems to be no reason for P-conscious states to be invoked. Since the 3T is primarily intended to test the robot's overall intelligence and linguistic understanding in the actual world, the A-conscious robot could conceivably pass the 3T while at the same time there *is nothing it is like* to be the 3T robot passing the test. We are now bordering on issues involved in demarcating the 'easy' from the 'hard' problems of consciousness, which, if pursued at this point, would be moving in a direction not immediately relevant to the topic at hand. So rather than exploring arguments relating to this deeper theme, I will simply contend that passing the 3T provides a sufficient condition for Block's version of A-consciousness, but not for P-consciousness, since it could presumably be passed by an artefact devoid of qualia.

Many critics of Block's basic type of view (including Searle [9] and Burge [10]) argue that if there can be such functional 'zombies' that are A-conscious but not P-conscious, then they are not genuinely conscious *at all*. Instead, A-consciousness is better characterized as a type of 'awareness', and is a form of consciousness only to the extent that it is parasitic upon P-conscious states. So we could potentially have a 3T for A-consciousness, but then the pivotal question arises, is A-consciousness without associated qualitative presentations really a form of *consciousness*? Again, I will not delve into this deeper and controversial issue in the present discussion, but simply maintain that the successful 3T robot does at least exhibit the type of A-*awareness* that people in, e.g., cognitive neuroscience tend to call consciousness. But as stated earlier, 'consciousness' is a multifaceted term, and there are also good reasons for *not* calling mere A-awareness without qualia a full-fledged form of consciousness.

For example, someone who was drugged or talking in their sleep could conceivably pass the 2T while still 'unconscious', that is A-'conscious' but not P-conscious. And a human sleep walker might even be able to pass the verbal and robotic 3T while 'unconscious' (again A-'conscious' but not P-conscious). What this seems to indicate is that only A-'consciousness' can be positively ascertained by behaviour. But there is an element of definitiveness here, since it seems plausible to say that an agent *could not* pass the 3T without being A-'conscious', at least in the minimal sense of A-awareness. If the robot were warned 'mind the banana peel' and it was *not* A-aware of the treacherous object in question on the ground before it, emitting the frequencies of electromagnetic radiation appropriate for 'banana-yellow', then it would not deliberately step over the object, but rather would slip and fall and fail the test.

5 A TOTAL TURING TEST FOR QUALIA

In the remainder of the paper I will not pursue the controversial issue as to whether associated P-consciousness is a necessary condition for concluding that the A-awareness of the successful 3T robot is genuinely a form of consciousness *at all*. Instead, I will explore an intensification of the standard 3T intended to prod more rigorously for *evidential support* of the presence of P-conscious states. This Total Turing Test for qualia (Q3T) is a

[2] Shieber [6] provides a valiant and intriguing rehabilitation/defense of the 2T, but it nonetheless still neglects crucial data, such as mastery of language exit and entry rules. Ultimately Shieber's rehabilitation in terms of interactive proof requires acceptance of the notion that *conversational* input/response patters alone are sufficient, which premise I would deny for the reasons given. The program is still operating within a closed syntactic bubble.

[3] See Schweizer [8] for an argument to the effect that even the combined linguistic and robotic 3T is still too weak as a definitive *behavioural* test of artificial intelligence.

more focused scrutiny of the successful 3T robot which emphasizes rigorous and extended verbal and descriptive probing into the qualitative aspects of the robot's purported internal experiences. So the Q3T involves unremitting questioning and verbal analysis of the robot's qualitative inner experiences, in reaction to a virtually limitless variety of salient external stimuli, such as paintings, sunsets, musical performances, tastes, textures, smells, pleasures and pains, emotive reactions...

Turing suggests a precursor version of this strategy in his 1950 discussion of the argument from consciousness, where he observes that the question of machine consciousness could be addressed by a sustained *viva voce*, where the artefact was asked questions directly concerning its aesthetic and other types of qualitative reactions and judgement in response to opened-ended questioning by the interrogator. Turing provides a conjectural illustration of the method in the guise of a 'sonnet writing' programme being quizzed by a human judge.

> Interrogator: In the first line of your sonnet which reads "Shall I compare thee to a summer's day," would not "a spring day" do as well or better?
>
> Witness: It wouldn't scan.
>
> Interrogator: How about "a winter's day," that would scan all right.
>
> Witness: Yes, but nobody wants to be compared to a winter's day.
>
> Interrogator: Would you say Mr. Pickwick reminded you of Christmas?
>
> Witness: In a way.
>
> Interrogator: Yet Christmas is a winter's day, and I do not think Mr. Pickwick would mind the comparison.
>
> Witness: I don't think you're serious. By a winter's day one means a typical winter's day, rather than a special one like Christmas.
>
> And so on....

The above sample dialogue serves as a good reminder of just how difficult the original 2T really is (and consequently why it hasn't yet been passed). However, this conjectured scenario is still confined to a merely conversational setting of verbal inputs and verbal outputs, and hence falls far short of the behavioural potential of a full 3T edition, as well as the concomitant evidential standards applicable in the human case.

Plebe and Perconti [11] put forward a strengthened adaptation of a 2T-style *viva voce*, where the pivotal difference is that, in addition to merely linguistic inputs, the computer must now give appropriate and testable conversational reactions to *uploaded images*. This is an interesting and important augmentation of the original 2T, since the inputs are no longer strictly linguistic, and the test is aimed at evaluating verbally plausible responses to stimuli that, *to us at least*, have a phenomenal aspect. As an example of the method, Plebe and Perconti supply an excerpt from a hypothetical conversation.

> Interrogator: Do you want to look at a picture of me?
>
> Machine: Oh yes, thanks, let's upload that.
> <.... uploading>
>
> Machine: Mmmh, I see several people here, who are you?
>
> Interrogator: Try to guess.
>
> Machine: Well, I know you are blond and shy, so I would guess the second from the left.

This appears to be an order of magnitude jump over the purely verbal 2T, and accordingly its standards of satisfaction are even more futuristic. However, in terms of the ultimate goal of providing a test, the passing of which constitutes a sufficient condition for the presence of genuine conscious experience in an artefact, it should be noted that the inputs, at a crucial level of analysis, remain purely syntactic and non-qualitative, in that the uploaded image must take the form of a digital file. Hence this could at most provide evidence of some sort of (proto) A-awareness in terms of salient data extraction and attendant linguistic conversion from a digital source, where the phenomenal aspects produced in humans by the original (pre-digitalized) image are systematically corroborated by the computer's linguistic outputs when responding to the inputted code.

Although a major step forward in terms of expanding the input repertoire under investigation, as well as possessing the virtue of being closer to the limits of practicality in the nearer term future, this proposed new qualia 2T still falls short of the full linguistic and robotic Q3T. In particular it tests, in a relatively limited manner, only one sensory modality, and in principle there is no reason why this method of scrutiny should be restricted to the intake of photographic images represented in digital form. Hence a natural progression would be to test a computer on uploaded audio files as well. However, this expanded 2T format is still essentially passive in nature, where the neat and tidy uploaded files are hand fed into the computer by the human interrogator, and the outputs are confined to mere verbal response. Active perception of and reaction to distal objects in the real world arena are critically absent from this test, and so it fails to provide anything like evidential parity with the human case. And given the fact that the selected non-linguistic inputs take the form of digitalized representations of possible visual (and/or auditory) stimuli, there is still no reason to think that there is anything it is like to be the 2T computer processing the uploaded encoding of an image of, say, a vivid red rose.

But elevated to a full 3T arena of shared external stimuli and attendant discussion and analysis, the positive evidence of a victorious computational artefact would become exceptionally strong indeed. So the extended Q3T is based on a methodology akin to Dennett's [12] 'heterophenomenology' - given the robot's presumed success at the standard Total Turing Test, we count this as behavioural evidence sufficient to warrant the application of the intentional stance, wherein the robot is treated as a rational agent harbouring beliefs, desires and various

other mental states exhibiting intentionality, and who's actions can be explained and predicted on the basis of the content of these states. Accordingly, the robot's salient sonic emissions are *interpreted* as natural language utterances asserting various propositions and expressing assorted contents. For the reasons delineated above in section 4, I would argue that this interpretive step and application of the intentional stance to a mere artefact is not evidentially warranted in a limited 2T type of setting, and hence in the latter case, the syntactic tokens emitted by the computer cannot be given the semantic value that we would attribute to utterances produced by our fellow human beings.

However, following Dennett, the gathering of scientifically rigorous empirical evidence proceeds with neutrality regarding the possible conscious experience of the robot. In this extended Q3T context, the artefact is queried not about uploaded images as 'qualitative' *input* data, but rather about *its own* subjective reactions and attendant introspective states. The robot is not hand fed selected digital files, but rather is presented with the same external stimuli as its human interlocutor, the latter of whom can then interrogate the robot regarding its purported qualitative experiences in response to the shared stimuli. Hence the human and the robot can both observe the same Dürer and the same Dali paintings, listen to the same Vivaldi and Schubert concertos, attend the same wine tasting event, eat the same brand of hamburger... And the robot must discourse on its subjective impressions and reactions like a true aesthete, and where this discourse can be calibrated by the human with respect to the latter's own qualitative reactions.

As an example of the method, an excerpt from a hypothetical Q3T interaction might look something like the following.

> Interrogator: Isn't the sunset over the Pacific magnificent?
>
> Robot: Yes, the colours are sublime. Let's go to the top of that rise for a better view...
> Excellent, now we have the full panorama.
>
> Interrogator: What impression does it make on you?
>
> Robot: As the glowing orange orb of the sun begins to touch the calm blue line of the horizon, yet another day is passing. Now the radiant disc begins to sink into the sea and the background crimson intensifies and grows warmer. It gives me a feeling of sadness, but at the same time a glimpse of ineffable beauty...
> Ouch!! Why the hell did you pinch me?
>
> Interrogator: Just wanted to bring you back down to earth - you were starting to wax a bit too lyrical.
>
> Robot: Thanks very much for your concern. Could you please pass me the bottle of delicious chilled Chardonnay - I want to savour another glass along with the last rays of the setting sun.
>
> Interrogator: Here you go.
>
> Robot: Arrrgh, that tastes disgusting! - what happened to the wine?
>
> Interrogator: Uhh, I just stirred in a little marmite when you weren't looking - wanted to see how you'd react. This is a Q3T, after all...

Even though a merely A-conscious robot could conceivably pass the verbal and robotic 3T while at the same time as there being *nothing it is like* for the robot passing the test, in this more focussed version of the 3T the robot would at least have to be able to go on at endless length *talking about* what it's like. And this talk must be in response to an open ended range of different combinations of sensory inputs, which are shared and monitored by the human judge. Such a test would be both subtle and extremely demanding, and it would be nothing short of remarkable if it could *not* detect a fake. And presumably a human sleepwalker who could pass a normal 3T as above would nonetheless *fail* this type of penetrating Q3T (or else wake up in the middle!), and it would be precisely on the grounds of such failure that we would infer that the human was actually asleep and not genuinely P-conscious of what was going on.

If sufficiently rigorous and extended, this would provide extremely powerful inductive evidence, and indeed to pass the Q3T the robot would have to attain full evidential parity with the human case, in terms of externally manifested behaviour.

6 BEYOND BEHAVIOUR

So on what grounds might one *consistently deny* qualitative states and P-consciousness in the case of the successful Q3T robot and yet grant it in the case of a behaviourally indistinguishable human? The two most plausible considerations that suggest themselves are both based on an appeal to essential differences of *internal* structure, either physical/physiological or functional/computational. Concerning the latter case, many versions of CTM focus solely on the functional analysis of propositional attitude states such as belief and desire, and simply ignore other aspects of the mind, most notably consciousness and qualitative experience. However others, such as Lycan [13], try to extend the reach of Strong AI and the computational paradigm, and contend that *conscious states* arise via the implementation of the appropriate computational formalism. Let us denote this extension of the basic CTM framework to the explanation of conscious experience 'CTM+'. And a specialized version of CTM+ might hold that qualitative experiences arise in virtue of the particular functional and information processing structure of the *human* brand of cognitive architecture, and hence that, even though the robot is indistinguishable in terms of input/output profiles, nonetheless its internal processing structure is sufficiently different from ours to block the inference to P-consciousness. So the non-identity of abstract functional or computational structure might be taken to undermine the claim that bare behavioural equivalence provides a sufficient condition for the presence of internal conscious phenomena.

At this juncture, the proponent of artificial consciousness might appeal to a version of Van Gulick's [14] defense of functionalism against assorted 'missing qualia' objections. When aimed against functionalism, the missing qualia arguments generally assume a deviant realization of the very same abstract computational procedures underlying human mental phenomena, in a world that's nomologically the same as

ours in all respects, and the position being supported is that consciousness is to be equated with states of the biological brain, rather than with any *arbitrary* physical state playing the same functional role as a conscious brain process. For example, in Block's [15] well known 'Chinese Nation' scenario, we are asked to imagine a case where each person in China plays the role of a neuron in the human brain and for some (rather brief) span of time the entire nation cooperates to implement the same computational procedures as a conscious human brain. The rather compelling 'common sense' conclusion is that even though the entire Chinese population may implement the same computational structure as a conscious brain, there are nonetheless no purely qualitative conscious states in this scenario outside the conscious Chinese *individuals* involved. And this is then taken as a counterexample to purely functionalist theories of consciousness.

Van Gulick's particular counter-strategy is to claim that the missing qualia argument begs the question at issue. How do we know, *a priori*, that the very same functional role *could* be played by arbitrary physical states that were unconscious? The anti-functionalist seems to beg the question by assuming that such deviant realizations are possible in the first place. At this point, the burden of proof may then rest on the functionalist to try and establish that there are in fact functional roles in the human cognitive system that could only be filled by *conscious* processing states. Indeed, this strategy seems more interesting than the more dogmatic functionalist line that isomorphism of abstract functional role *alone* guarantees the consciousness of any physical state that happens to implement it.

So to pursue this strategy, Van Gulick examines the psychological roles played by phenomenal states in humans and identifies various cognitive abilities which *seem* to require both conscious and self-conscious awareness – e.g. abilities which involve reflexive and meta-cognitive levels of representation. These include things like planning a future course of action, control of plan execution, acquiring new non-habitual task behaviours. These and related features of human psychological organization seem to require a conscious self-model. In this manner, conscious experience appears to play a *unique* functional role in broadcasting 'semantically transparent' information throughout the brain. In turn, the proponent of artificial consciousness might plausibly claim that the successful Q3T robot must possesses analogous processing structures in order to evince the equivalent behavioural profiles when passing the test. So even though the processing structure might not be identical to that of human cognitive architecture, it must nonetheless have the same basic cognitive abilities as humans in order to pass the Q3T, and if these processing roles in humans require phenomenal states, then the robot must enjoy them as well.

However, it is relevant to note that Van Gulick's analysis seems to blur Block's distinction between P-consciousness and A-consciousness, and an obvious rejoinder at this point would be that all of the above processing roles in both humans and robots could in principle take place with only the latter and not the former. Even meta-cognitive and 'conscious' self models could be accounted for merely in terms of A-*awareness*. And this brings us back to the same claim as in the standard 3T scenario - that even the success of the Q3T robot could conceivably be explained without invoking P-consciousness *per se*, and so it still fails as a sufficient condition for attributing full blown qualia to computational artefacts.

7 MATTER AND CONSCIOUSNESS

Hence functional/computational considerations seem too weak to ground a positive conclusion, and this naturally leads to the question of the physical/physiological status of qualia. If even meta-cognitive and 'conscious' self models in humans could in principle be accounted for merely in terms of A-*awareness*, then how and why do *humans* have purely qualitative experience? One possible answer could be that P-conscious states are essentially *physically based phenomena*, and hence result from or supervene upon the particular structure and causal powers of the actual central nervous system. And this perspective is re-enforced by what I would argue (on the following independent grounds) is the fundamental inability of abstract functional role to provide an adequate theoretical foundation for qualitative experience.

Unlike computational formalisms, conscious states are inherently *non-abstract*; they are *actual*, occurrent phenomena extended in physical time. Given multiple realizability as a hallmark of the theory, CTM+ is committed to the result that qualitatively identical conscious states are maintained across widely different kinds of physical realization. And this is tantamount to the claim that an actual, substantive and *invariant* qualitative phenomenon is preserved over radically diverse real systems, while at the same time, *no* internal physical regularities need to be preserved. But then there is no actual, occurrent factor which could serve as the causal substrate or supervenience base for the substantive and invariant phenomenon of internal conscious experience. The advocate of CTM+ cannot rejoin that it is *formal role* which supplies this basis, since formal role is abstract, and such abstract features can only be *instantiated* via actual properties, but they do not have the power to *produce* them.

The only (possible) non-abstract effects that instantiated formalisms are required to preserve must be specified in terms of their input/output profiles, and thus *internal* experiences, qua actual events, are in principle omitted. So (as I've also been argued elsewhere: see Schweizer [16,17]) it would appear that the non-abstract, occurrent nature of conscious states entails that they must depend upon intrinsic properties of the brain as a proper subsystem of the actual world (on the crucial assumption of *physicalism* as one's basic metaphysical stance - obviously other choices, such as some variety of dualism, are theoretical alternatives). It is worth noting that from this it *does not follow* that other types of physical subsystem could not share the relevant intrinsic properties and hence also support conscious states. It only follows that they would have this power in virtue of their intrinsic physical properties and *not* in virtue of being interpretable as implementing the same abstract computational procedure.

8 CONCLUSION

We know by direct first person access that the human central nervous system is capable of sustaining the rich and varied field of qualitative presentations associated with our normal cognitive activities. And it certainly *seems as if* these presentations play a

vital role in our mental lives. However, given the above critical observation regarding Van Gulick's position, *viz*., that all of the salient processing roles in *both* humans and robots could in principle take place strictly in terms of A-awareness without P-consciousness, it seems that P-conscious states are not actually necessary for explaining observable human behaviour and the attendant cognitive processes. In this respect, qualia are rendered *functionally* epiphenomenal, since purely qualitative states per se are not strictly required for a functional/computational account of human mentality. However, this is not to say that they are *physically* epiphenomenal as well, since it doesn't thereby follow that this aspect of physical/physiological structure does not in fact play a causal role in the particular *human* implementation of this functional cognitive architecture. Hence it becomes a purely contingent truth that humans have associated P-conscious experience.

And this should not be too surprising a conclusion, on the view that the human mind is the product of a long course of exceedingly happenstance biological evolution. On such a view, perhaps natural selection has simply *recruited* this available biological resource to play vital functional roles, which in principle could have instead been played by P-*unconscious* but A-aware states in a *different type* of realization. And in this case, P-conscious states in humans are thus a form of 'phenomenal overkill', and nature has simply been an opportunist in exploiting biological vehicles that happened to be on hand, to play a role that could have been played by a more streamlined and less rich type of state, but where a 'cheaper' alternative was simply not available at the critical point in time. Evolution and natural selection are severely curtailed in this respect, since the basic ingredients and materials available to work with are a result of random mutation on existing precursor structures present in the organism(s) in question. And perhaps human computer scientists and engineers, not limited by what happens to get thrown up by random genetic mutations, have designed the successful Q3T robot utilizing a cheaper, artificial alternative to the overly rich biological structures sustained in humans.

So in the case of the robot, it would remain an open question whether or not the physical substrate underlying the artefact's cognitive processes had the requisite causal powers or intrinsic natural characteristics to sustain P-conscious states. Mere behavioural evidence on its own would not be sufficient to adjudicate, and an independent standard or criterion would be required.[4] So if P-conscious states are thought to be essentially physically based, for the reasons given above, and if the robot's Q3T success could in principle be explained through appeal to mere A-aware stets on their own, then it follows that the non-identity of the artefact's physical structure would allow one to consistently extend Turing's polite convention to one's conspecifics and yet withhold it from the Q3T robot.

[4] This highlights one of the intrinsic limitations of the Turing test approach to such questions, since the test is designed as an *imitation game*, and humans are the ersatz target. Hence the Q3T robot is designed to behave as if it had subjective, qualitative inner experiences indistinguishable from those of a human. However, if human qualia are the products of our particular internal structure (either physical-physiological or functional-computational), and if the robot is significantly different in this respect, then the possibility is open that the robot might be P-conscious and yet fail the test, simply because its resulting qualitative experiences are significantly different than ours. And indeed, a possibility in the reverse direction is that the robot might even *pass* the test and sustain an entirely different phenomenology, but where this internal difference is not manifested in its external behaviour.

REFERENCES

[1] A. Turing, 'Computing machinery and intelligence', *Mind* 59: 433-460 (1950).

[2] N. Block, 'Psychologism and behaviorism', *Philosophical Review* 90: 5-43 (1981).

[3] R. French, 'The Turing test: the first 50 years', *Trends in Cognitive Sciences* 4: 115-122 (2000).

[4] N. Block, 'On a confusion about a function of consciousness', *Behavioral and Brain Sciences* 18, 227-247, (1995).

[5] B. Baars, *A Cognitive Theory of Consciousness*, Cambridge University Press, (1988).

[6] S. Shieber, 'The Turing test as interactive proof', *Nous* 41:33-60 (2007).

[7] S. Harnad 'Other bodies, other minds: A machine incarnation of an old philosophical problem', *Minds and Machines* 1: 43-54, (1991).

[8] P. Schweizer, 'The externalist foundations of a truly total Turing test', *Mind & Machines*, DOI 10.1007/s11023-012-9272-4, (2012).

[9] J. Searle, *The Rediscovery of the Mind*, MIT Press, (1992).

[10] T. Burge, 'Two kinds of consciousness', in N. Block et al. (eds), *The Nature of Consciousness: Philosophical Debates*, MIT Press, (1997).

[11] A. Plebe and P. Perconti, 'Qualia Turing test: Designing a test for the phenomenal mind', in Proceedings of the First International Symposium *Towards a Comprehensive Intelligence Test (TCIT), Reconsidering the Turing Test for the 21st Century*, 16-19, (2010).

[12] D. Dennett, *Consciousness Explained*, Back Bay Books, (1992).

[13] W. G., Lycan, *Consciousness*, MIT Press, (1987).

[14] R. Van Gulick, 'Understanding the phenomenal mind: Are we all just armadillos?', in *Consciousness: Psychological and Philosophical Essays*, M. Davies and G. Humphreys (eds.), Blackwell, (1993).

[15] N. Block, 'Troubles with functionalism', in C. W. Savage (ed), *Perception and Cognition*, University of Minnesota Press, (1978).

[16] P. Schweizer, 'Consciousness and computation.' *Minds and Machines*, 12, 143-144, (2002)

[17] P. Schweizer, 'Physical instantiation and the propositional attitudes', *Cognitive Computation*, DOI 10.1007/s12559-012-9134-7, (2012).